Enhancing an English-Polish electronic dictionary for multiword expression research

Piotr Bański
Institute of English Studies,
University of Warsaw
E-mail: bansp@o2.pl

Radosław Moszczyński
Department of Formal Linguistics,
University of Warsaw
E-mail: r.moszczynski@uw.edu.pl

Aims of the project

•Cleaning up the existing markup in a large, electronic English-Polish dictionary and preparing a TEI P5 customization for it, aiming at full P5 conformance within the TEI namespace.
•Recoding the legacy plain-text IDAREX (see below) multiword-expression formulas as chunks of XML, using a small RELAX NG regular expression grammar. The new format will extend the plain-text-based IDAREX formalism beyond its current limitations.
•Establishing methods of visualizing the formulas within dictionary entries together with methods allowing for enhanced navigation, e.g. across structures of the same syntactic type.

The dictionary

The project is based on an English-Polish dictionary compiled by Tadeusz Piotrowski and Zygmunt Saloni. The dictionary was originally a traditional publication that appeared in print as Piotrowski and Saloni (1992). Its initial electronic format was pure TeX. In the late 1990s, it was converted into SGML and its idiom definitions were enriched with the IDAREX (IDioms As REgular eXpressions) formalism (the conversion process is described in Piotrowski, 1999). Several years later, an XML version of the dictionary has been created, encoded in early TEI P5, and recently the dictionary has been released by the authors as TeX, SGML and XML source, dual-licensed under GNU GPL 2 and GNU FDL 1.2.

Brief introduction to IDAREX

For the purposes of the STEEL (Developing Specialized Translation/Foreign Language Understanding Tools for Eastern European Languages) project, all idioms in the original dictionary by Piotrowski and Saloni had to be encoded in IDAREX, a formalism developed by Xerox (see Breidt, Segond & Valetto, 1996). Below is the idiom *to take the bull by the horns* encoded in IDAREX:

```
ADV* take Verb: :the :bull :by :the :horns
```

A prepended colon marks a surface form, and a postpended colon marks what in the Two-Level Morphology framework of Kimmo Koskenniemi is called the lexical level. Surface level tokens are unchangeable, while words at the lexical level can be inflected. ADV stands for any adverb, and * means it can appear any number of times. See the full article for more information on IDAREX.

Why recode IDAREX as XML?

- Representation: instead of strings that for each parse would have to be tokenized in a non-trivial way (cf. “take Verb:”, which is a single token), the XML representation makes it possible to encode some structural aspects of the IDAREX (by means of the <choice> or <zeroOrMore> elements).
- Atomization: each element of the XML representation can be furnished with the part of speech that it represents and with a hyperlink to the entry/sense where it is described.
- The next logical step in the research is extracting common patterns from XML representations (at this stage, the IDAREX descriptions may be moved to a separate file, cross-linked from the dictionary); this means abstracting from concrete tokens toward part-of-speech variables.
- Armed with this kind of representations, one may attempt automatic identification of idioms in corpora.

From IDAREX to RelaxNG

This section presents some examples of moving from IDAREX to a RelaxNG representation. In the following expression, “ADV*” signals that an adverb may be used any number of times, and “take Verb:” is used to restrict the word *take* to its verbal sense and to signal that it can appear in any of its inflectional forms. Note that the token “take Verb:” is recoded as a single XML element with an appropriate @pos attribute. Note also that in our formalism, the @pos attributes of variables and lexical tokens are uniform, which is a basis for further applications (e.g. the establishment of possible references and substitutions). Because we want to proceed step-by-step, the @pos attributes for <surface> elements are for the time being optional. The same is true of @entry attributes which (after lemmatization) will point at the relevant entries of the dictionary.

```
ADV* take Verb: :the :bull :by :the :horns
```

```
<sequence>
  <zeroOrMore>
    <var>ADV</var>
  </zeroOrMore>
  <lexical pos="V">take</lexical>
  <surface>the</surface>
  <surface>bull</surface>
  <surface>by</surface>
  <surface>the</surface>
  <surface>horns</surface>
</sequence>
```

The second example is a formula requiring a choice of one out of two related sequences. This time the XML is pasted from an actual fragment of the dictionary, where it is located in the <gramGrp> element, storing grammatical information. The dictionary with such extensions is valid against a version of TEI P5 1.0 Guidelines with our additional IDAREX module plugged in.
An additional virtue of this representation is the possibility of relating phrasal and word-level categories by means of the @pos and @extent attributes. Something that we want to implement soon is a possibility of relating elements of such alternative sequences (notice that the two NP symbols in the IDAREX expression designate the same object). The XPointer element() function seems an excellent tool for this purpose.

```
[take V: NP :into :account |
take V: :account :of NP]
```

```
<gramGrp>
  <idx:idarex>
    <idx:choice>
      <idx:sequence>
        <idx:lexical pos="v">take
        </idx:lexical>
        <idx:var pos="n"
extent="phrase">NP</idx:var>
        <idx:surface>into</idx:surface>
        <idx:surface>account</idx:surface>
      </idx:sequence>
      <idx:sequence>
        <idx:lexical pos="v">take</idx:lexical>
        <idx:surface>account</idx:surface>
        <idx:surface>of</idx:surface>
        <idx:var pos="n"
extent="phrase">NP</idx:var>
      </idx:sequence>
    </idx:choice>
  </idx:idarex>
</gramGrp>
```

At this stage of the project, in our test dictionary, we keep the XML-ized IDAREX specifications inside entries, as content of the appropriate `re/gramGrp/gram` elements, in order to keep them where they belong, according to their lexical and semantic relationships. At a later stage of the project, we will consider externalizing the IDAREX level of annotation to a separate file.

References

Breidt, Elisabeth; Segond, Frédérique; Valetto, Giuseppe. (1996). Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of the 16th Conference on Computational Linguistics*, Volume 2.Morristown, NJ: Association for Computational Linguistics, pp. 1036–1040.

Piotrowski, Tadeusz. (1999). Tagging and Conversion of a Bilingual Dictionary for XeLDA, a Xerox Computer Assisted Translation System. In *Papers in Computational Lexicography COMPLEX '99 Proceedings*. Budapest: Hungarian Academy of Sciences, pp. 113-120.

Piotrowski, Tadeusz & Saloni, Zygmunt. (1992). *Nowy słownik angielsko-polski polsko-angielski [New English-Polish, Polish-English dictionary]*. Warszawa: Editions Spotkania.